8С

УДК 519.6:004.93

С.А. Субботин, А.А. Олейник
Запорожский национальный технический университет, Украина
subbotin@zntu.edu.ua

Выделение набора информативных признаков на основе эволюционного поиска с кластеризацией

Решается задача поиска наиболее информативной комбинации признаков с помощью методов эволюционной оптимизации. Предложен метод эволюционного поиска с кластеризацией признаков. Проведены эксперименты по выделению информативного набора признаков для синтеза моделей классификации автотранспортных средств.

Введение

В автоматизированных системах распознавания образов, принятия решений, прогнозирования и классификации важное значение имеет описание объекта на основе набора (системы) признаков, обладающего максимальной информативностью. Использование неинформативных и избыточных признаков не только оказывается бесполезным, но и снижает эффективность процесса распознавания [1]. Поэтому при синтезе распознающих моделей актуальным является этап поиска наиболее значимой комбинации признаков.

Задача выделения наиболее значимого признакового набора из исходного множества данных заключается в поиске такой комбинации информативных признаков, при которой достигается минимум заданного критерия оценивания набора признаков.

В настоящее время известны различные методы отбора признаков [2], [3]: метод полного перебора, эвристические методы последовательного добавления и удаления признаков, ранжирование признаков. Однако такие методы связаны с необходимостью комбинаторного перебора, что делает их мало применимыми на практике, либо используют критерии оценивания индивидуальной информативности признаков, не учитывая при этом совместное влияние всего набора признаков на выходной параметр.

Для выделения наиболее значимой комбинации признаков могут быть использованы методы эволюционного поиска [4-6], которые на каждой итерации работают не с единственным решением, а с некоторым множеством решений, что позволяет во многих случаях анализировать пространство поиска быстрее по сравнению с традиционными методами, не выдвигая при этом дополнительных требований к виду целевой функции, что, в свою очередь, приводит к более быстрому поиску оптимума.

Однако при отборе признаков с помощью классических методов эволюционной оптимизации [4-6] не учитывается расположение признаков в пространстве экземпляров, в результате чего новые решения, генерируемые в процессе поиска, могут включать в себя малоинформативные признаки, что приводит к формированию и оцениванию изначально неинформативных наборов.

Целью настоящей работы является создание метода эволюционного поиска с кластеризацией признаков, который учитывает расположение признаков в пространстве экземпляров при формировании новых решений и позволяет выделить комбинацию информативных признаков, принадлежащих разным факторным группам.

Отбор информативных признаков на основе методов эволюционной оптимизации

Использование эволюционного поиска требует определения способа представления информации в хромосоме и фитнесс-функции, с помощью которой производится оценивание эффективности хромосом для решаемой задачи.

Для отбора информативных признаков из исходного массива, содержащего L признаков, с помощью методов эволюционного поиска решение (хромосома) представляется битовой строкой размера L. Если бит хромосомы принимает единичное значение, тогда соответствующий ему признак считается информативным и учитывается при оценивании набора признаков, соответствующего хромосоме. В противном случае, когда бит принимает нулевое значение, признак считается неинформативным и не используется при оценке комбинации признаков.

Преимущество такого представления заключается в том, что классические эволюционные операторы скрещивания и мутации могут быть применены для отбора признаков без каких-либо изменений.

При поиске наиболее значимой комбинации признаков на основе методов эволюционной оптимизации в качестве фитнесс-функции хромосом используются критерии, позволяющие оценить информативность набора признаков, соответствующего оцениваемой хромосоме.

В качестве таких критериев используются: показатели эффективности классификации или прогнозирования по моделям, синтезированным на основе оцениваемых комбинаций признаков либо фильтрующие критерии.

Критерии, относящиеся к первой группе, оценивают набор признаков с помощью ошибки прогнозирования или классификации по модели, построенной на основе признаков из анализируемого набора. В качестве синтезируемых моделей могут использоваться регрессионные, нечеткие, нейросетевые, нейро-нечеткие и другие.

В случае отбора признаков при решении задачи прогнозирования в качестве критериев оценивания информативности могут быть использованы: среднеквадратическая ошибка, сумма квадратов отклонений, средняя абсолютная ошибка, сумма значений абсолютных отклонений, максимальное абсолютное отклонение, средняя относительная ошибка, сумма относительных отклонений, максимальное относительное отклонение [7], [8]. При отборе признаков для классификации используются: вероятность принятия ошибочных решений и критерий Фишера [1], [3], [7], [8].

Наиболее часто для оценивания информативности набора признаков X используется сумма квадратов отклонений: $E(X)=\sum_{p=1}^{m}\left(y^p-y_X^p\right)^2$, где $E(X)$ – сумма квадратов отклонений реальных значений выходного параметра от значений, вычисленных с помощью модели, синтезированной на основе комбинации признаков X; $y^p{}_X$ – значение выходного параметра p-го экземпляра, рассчитанное по синтезированной модели; y^p – реальное значение выходного параметра p-го экземпляра; m – количество экземпляров в исходной выборке данных.

Использование ошибок синтезируемых моделей для оценивания информативности набора признаков является достаточно ресурсоемкой процедурой, поскольку синтез моделей на основе оцениваемой комбинации признаков занимает значительно большее время по сравнению с оцениванием признаков путем применения фильтрующих критериев оценивания совместного влияния признаков.

Как правило, использование таких критериев приводит к лучшим результатам по сравнению с фильтрующими критериями, поскольку они ориентированны на поиск информативной комбинации признаков для конкретной модели, которая в дальнейшем будет применяться на практике.

Однако это приводит к уменьшению гибкости результатов в виде набора информативных признаков. И в случае принятия решения об изменении типа модели, используемой для описания исследуемого объекта или процесса, необходимо будет запускать метод для повторного поиска комбинации информативных признаков, соответствующей новой модели.

На практике часто возникают ситуации, когда исходный набор данных содержит чрезвычайно большие объёмы информации. Это приводит к значительным вычислительным и временным затратам на построение модели, описывающей исследуемый объект или процесс. В результате построение моделей и использование их ошибок для оценивания хромосом является неприемлемым.

В таких случаях используют фильтрующие критерии, которые предполагают исключение неинформативных признаков из исходного набора до построения модели, описывающей исследуемый объект.

Одним из преимуществ таких методов оценивания информативности является то, что они не нуждаются в повторном запуске в случае необходимости синтеза новой модели по уже отобранным признакам. Фильтры являются вычислительно более простыми по сравнению с другими критериями и эффективно могут применяться для отбора информативных признаков из массивов данных очень большого размера.

Однако в результате использования фильтрующих критериев могут быть получены такие комбинации признаков, на основе которых не удастся построить модель, обеспечивающую требуемую точность. Это вызвано тем, что такие критерии непосредственно не связаны с математической моделью, которая будет использоваться для описания исследуемого объекта.

К фильтрующим критериям, используемым для оценивания признаков, могут быть отнесены: множественный коэффициент корреляции, коэффициент корреляции Пирсона, дисперсионное отношение [9], коэффициент связи [10], информационный критерий, энтропия набора признаков [1], критерий, основанный на статистическом подходе [2].

Упомянутые критерии оценивания информативности комбинаций признаков не учитывают количество отобранных признаков. Поэтому в качестве фитнесс-функции предлагается использовать выражение, минимизирующее количество отобранных признаков и критерий оценивания информативности набора признаков:

$$J(H_j) = \left(1 + \frac{1}{L}\sum_{i=1}^{L} h_{ij}\right) I_j ,$$

где I_j – критерий оценивания совместного влияния набора признаков, соответствующего оцениваемой хромосоме H_j.

Предложенный критерий позволит обеспечить эффективное оценивание хромосом с учетом информативности оцениваемой комбинации и количества признаков, содержащихся в ней.

Эволюционный поиск с кластеризацией признаков

Существенным недостатком классических методов эволюционного поиска при отборе признаков является то, что они не учитывают близости расположения признаков в пространстве экземпляров, в результате чего новые комбинации признаков (хромосомы), формируемые путем применения эволюционных операторов инициализации, скрещивания и мутации, могут включать в себя признаки, содержащие одинаковую информацию об исследуемом объекте. Очевидно, что наборы признаков, соответствующие таким хромосомам, являются малоинформативными или избыточными.

В разработанном методе эволюционного поиска с кластеризацией признаков предлагается группировать схожие признаки с помощью методов кластеризации, которые позволяют разбить выборку на группы компактно расположенных признаков в пространстве экземпляров (кластеры, факторные группы) и выделить в каждом кластере по одному наиболее типичному признаку.

При формировании новых хромосом в результате применения операторов инициализации, скрещивания и мутации предлагается рассчитывать вероятность включения признака в хромосому, которая зависит от расположения признака в кластере (расстояния от него до центра кластера), индивидуальной информативности признака, а также индивидуальной информативности центра его кластера.

Эволюционный поиск с кластеризацией признаков для выделения наиболее значимого набора признаков из заданной выборки $<X, Y>$ предлагается выполнять как следующую последовательность шагов.

Шаг 1. Сгруппировать признаки исходной выборки данных в кластеры.

Шаг 1.1. Для каждого признака X_i рассчитать Эвклидово расстояние от него до всех остальных признаков в выборке. Эвклидово расстояние между признаками X_a и X_b вычисляется по формуле: $d_E(X_a; X_b) = \sqrt{\sum_{p=1}^{m}(x_{pa} - x_{pb})^2}$, где m – количество экземпляров в выборке; x_{pa} и x_{pb} – значения a-го и b-го признаков p-го экземпляра соответственно.

Шаг 1.2. На основе рассчитанных ранее расстояний между экземплярами, используя методы кластер-анализа [2], [11], например, метод с добавлением кластеров, метод с удалением кластеров, комбинированный метод или метод нечетких C-средних, сформировать группы признаков, компактно расположенных в пространстве экземпляров. Выделить признаки, являющиеся центрами кластеров.

Шаг 1.3. Для каждого признака X_i вычислить вероятность его включения в хромосому.

Шаг 1.3.1. Рассчитать значение индивидуальной оценки информативности I_i признака X_i, например, на основе коэффициента парной корреляции, коэффициента корреляции знаков, коэффициента корреляции Фехнера, дисперсионного отношения, коэффициента связи, информационного критерия, энтропии признака, критерия, основанного на вероятностном подходе, или критерия, основанного на статистическом подходе [7], [8].

Шаг 1.3.2. Определить вероятность P_i включения i-го признака в хромосому:

$$P_i = I_i + \frac{d_E(X_i; X_{c,i})}{d_{E\max,c}}(I_i - I_c),$$

где $d_E(X_i; X_{c,i})$ – расстояние от признака X_i до центра его кластера; $d_{E\max,c}$ – максимальное расстояние в кластере, в котором расположен i-ый признак; I_c – информативность признака, являющегося центром кластера, в котором расположен признак X_i.

Шаг 2. Установить счетчик итераций (времени): $t = 0$.

Шаг 3. Инициализировать начальную популяцию из N хромосом, состоящих из L генов.

Шаг 3.1. Установить счетчик сформированных хромосом: $j = 1$.

Шаг 3.2. Сформировать j-ую хромосому H_j.

Шаг 3.2.1. Установить счетчик определенных генов: $i = 1$.

Шаг 3.2.2. Сгенерировать случайное число: $r = \mathrm{rand}[0;1]$, где $\mathrm{rand}[a; b]$ – случайно сгенерированное число в интервале $[a; b]$.

Шаг 3.2.3. Если $P_i > r$, тогда i-му гену j-ой хромосомы присвоить значение: $h_{ij} = 1$, в противном случае: $h_{ij} = 0$.

Шаг 3.2.4. Если j-ая хромосома сформирована полностью ($i = L$), тогда выполнить переход к шагу 3.3.

Шаг 3.2.5. Установить: $i = i + 1$.

Шаг 3.2.6. Перейти к шагу 3.2.2.

Шаг 3.3. Если сформированы все хромосомы ($j = N$), тогда выполнить переход к выполнению шага 4.

Шаг 3.4. Установить: $j = j + 1$.

Шаг 3.5. Перейти к шагу 3.2.

Шаг 4. Вычислить значение фитнесс-функции $f(H_j)$ хромосом текущей популяции по формуле:

$$f(H_j) = \frac{J(H_j)\sum_{i=1}^{L} h_{ij}}{\left(1 + \sum_{i=1}^{L} I_i h_{ij}\right)\left(1 + \sum_{i=1}^{L} P_i h_{ij}\right)},$$

где $J(H_j)$ – значение критерия, учитывающего размер и информативность набора признаков, соответствующего хромосоме H_j.

Шаг 5. Выполнить проверку критериев останова (достижение максимально допустимого времени функционирования метода, числа итераций, значения фитнесс-функции). Если критерии окончания поиска удовлетворены, тогда выполнить переход к шагу 11.

Шаг 6. Увеличить счетчик итераций: $t = t + 1$.

Шаг 7. Выбрать хромосомы для скрещивания и мутации путем использования одного из существующих методов отбора (пропорциональный отбор, отбор с использованием рулетки, турнирный отбор, пороговый отбор, отбор ранжированием).

Шаг 8. Применить оператор равномерного скрещивания. При этом в маске скрещивания установить единичные значения для генов, которым соответствуют признаки с вероятностью включения в хромосому, выше средней, остальным генам присвоить нулевые значения.

Шаг 9. Применить оператор точечной мутации. Вероятность мутации P_{Mi} i-го гена в мутирующей хромосоме предлагается рассчитывать по формуле: $P_{Mi} = \alpha\,(1 - P_i)$, где α – коэффициент, определяющий степени мутации, $\alpha \in [0; 1]$.

Шаг 10. Сформировать новое поколение. Выполнить переход к шагу 4.

Шаг 11. Останов.

Таким образом, в предложенном методе эволюционного поиска с кластеризацией признаков учитывается близость расположения признаков в пространстве экземпляров, что позволяет формировать новые решения из признаков, расположенных, как правило, в разных группах, увеличивая вероятность отыскания комбинации признаков, обладающей максимальной информативностью.

Эксперименты и результаты

Разработанный метод эволюционного поиска с кластеризацией признаков был программно реализован на языке пакета Matlab. Для проверки эффективности применения предложенного метода и разработанного программного обеспечения решалась задача отбора информативных признаков для синтеза распознающих моделей автотранспортных средств [12].

Исходная выборка, предоставленная ООО «МПА Групп», содержала преобразованные графические изображения различных транспортных средств, полученные с камер наблюдения. Выборка состояла из 1062 экземпляров, каждый из которых характеризовался 4096 признаками, представляющими собой нормированные значения интенсивности точек изображения, спроецированного на сенсорную матрицу точек размерностью 64×64, по которым определялись значения расчетных (искусственных) 26 признаков, обобщающих графическую информацию об объекте: x_1 – высота региона интереса; x_2 – ширина региона интереса; x_3 – коэффициент горизонтальной симметрии региона интереса; x_4 – коэффициент вертикальной симметрии региона интереса; x_5 – максимальное значение яркости области точек; x_6 – минимальное значение яркости области точек; x_7 – усредненное значение яркости области точек; x_8 – центральный момент второго порядка (дисперсия) гистограммы яркости области точек; x_9 – ассиметрия гистограммы яркости области точек; x_{10} – эксцесс гистограммы яркости области точек; x_{11} – нормированный дескриптор относительной гладкости; x_{12} – однородность; x_{13} – взвешенная однородность; x_{14} – средняя энтропия; x_{15} – максимальный модуль градиента яркости точек; x_{16} – максимальный модуль градиента со знаком; x_{17} – межпиксельная контрастность; x_{18} – бета-коэффициент; x_{19} – упрощенный бета-коэффициент; x_{20} – нормированный центральный момент двумерной функции яркости области точек; x_{21} – x_{26} – инвариантные моменты двумерной функции яркости области точек $I_i^1 - I_i^6$.

Выделение комбинации признаков выполнялось на основе методов эволюционного поиска. Начальные значения параметров эволюционных методов устанавливались следующими: оператор отбора – использованием рулетки, оператор скрещивания – равномерный, оператор мутации – точечный, количество особей в популяции $N = 100$, вероятность скрещивания $p_{скр} = 0{,}8$, вероятность мутации $p_{m} = 0{,}05$, максимальное количество итераций $T = 100$, количество элитных особей $N_e = 2$.

В качестве критерия оценивания информативности набора признаков использовалась среднеквадратическая ошибка классификации по двухслойной нейросети прямого распространения, синтезированной на основе признаков оцениваемой хромосомы и содержащей 4 нейрона на первом слое и один нейрон на втором слое. Все нейроны имели сигмоидную функцию активации, а в качестве дискриминантных функций использовалась взвешенная сумма.

Результаты экспериментов приведены в табл. 1, где t – время, затраченное на эволюционный поиск комбинации информативных признаков, сек.; k_f – количество вычисленных значений фитнесс-функции; k – количество отобранных признаков; ε – достигнутая ошибка прогнозирования.

Таблица 1 – Результаты отбора признаков с помощью различных эволюционных методов

Метод отбора признаков	Критерий			
	t	k_f	k	ε
Классический эволюционный поиск	183,9	122	10	0,0075
Эволюционный поиск с кластеризацией признаков	113,8	81	9	0,0062

Из табл. 1 видно, что при использовании предложенного метода эволюционного поиска с кластеризацией признаков для выделения информативной комбинации признаков затрачивается меньше времени и требуется меньше обращений к целевой функции по сравнению с классическим эволюционным поиском. При этом оптимальный набор содержит меньшее количество признаков и позволяет синтезировать модель (табл. 2), обеспечивающую лучшую точность классификации.

Таблица 2 – Матрица весовых коэффициентов нейростевой модели

Номер слоя	Номер нейрона в слое	Номер входа нейрона									
		0	1	2	3	4	5	6	7	8	9
1	1	–14,57	32,96	10,71	–8,418	7,729	0,924	–11,37	–3,822	–7,538	5,823
	2	–15,633	28,02	6,739	–26,81	–31,57	–29,62	38,773	15,638	31,372	2,616
	3	6,5428	–9,43	18,78	–15,51	4,193	16,625	16,74	–5,725	–23,85	3,824
	4	–2,637	–4,36	–26,53	9,725	–6,714	–0,842	–5,882	18,621	29,629	–7,518
2	1	82,747	–78,53	–54,72	–49,42	–43,56					

Заключение

В работе решена задача отбора информативных признаков для синтеза эффективных моделей исследуемых объектов, процессов и систем на основе эволюционного подхода.

Научная новизна работы заключается в том, что разработан метод эволюционного поиска с кластеризацией признаков, в котором учитывается близость расположения признаков в пространстве экземпляров. Это позволяет формировать новые решения из признаков, расположенных, как правило, в разных группах, увеличивая вероятность отыскания комбинации признаков, обладающей наибольшей значимостью.

Практическая ценность результатов работы состоит в том, что разработано программное обеспечение, реализующее предложенный метод отбора признаков, а также решена задача выделения информативного набора признаков для синтеза моделей классификации автотранспортных средств.

Исследование выполнено в рамках НИР «Научно-методические основы и математическое обеспечение для автоматизации и моделирования процессов управления и поддержки принятия решений на основе процедур распознавания и эволюционной опти-

мизации в нейросетевом и нечетклогическом базисах» (№ гос. регистрации 0106U008621) и «Разработка методов и программных средств на основе обучения, распознавания, оптимизации и адаптации для принятия решений в автоматизированных системах управления транспортными средствами» (№ гос. регистрации 0107U0006781).

Литература

1. Биргер И.А. Техническая диагностика. – М.: Машиностроение, 1978. – 240 с.
2. Интеллектуальные средства диагностики и прогнозирования надежности авиадвигателей: Монография / В.И. Дубровин, С.А. Субботин, А.В. Богуслаев, В.К. Яценко. – Запорожье: ОАО «Мотор-Сич», 2003. – 279 с.
3. Прикладная статистика: Классификация и снижение размерности: Справ. изд. / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин. – М.: Финансы и статистика, 1989. – 607 с.
4. Haupt R., Haupt S. Practical Genetic Algorithms. – New Jersey: John Wiley & Sons, 2004. – 261 p.
5. Субботин С.А., Олейник А.А. Ускоренный метод эволюционного отбора признаков // Автоматика-2006: Тези доповідей Тринадцятої Міжнародної науково-технічної конференції (25 – 28 вересня 2006 р.). – Вінниця: УНІВЕРСУМ – Вінниця, 2006. – С. 409.
6. Subbotin S., Oleynik A. Entropy Based Evolutionary Search for Feature Selection // The experience of designing and application of CAD systems in Microelectronics: Proceedings of the IX International Conference CADSM-2007 (20 – 24 February 2007). – Lviv: Publishing house of Lviv Polytechnic, 2007. – P. 442-443.
7. Guyon I., Elisseeff A. An Introduction to Variable and Feature Selection // Journal of Machine Learning Research. – 2003. – № 3. – P. 1157-1182.
8. Dash M., Liu H. Feature Selection for Classification // Intelligent Data Analysis. – 1997. – № 1. – P. 131-156.
9. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Исследование зависимостей. – М.: Финансы и статистика, 1985. – 487 с.
10. Дисперсионная идентификация / Под ред. Н.С. Райбмана. – М.: Наука, 1981. – 336 с.
11. Классификация и кластер / Под ред. Дж. Вэн Райзина. – М.: Мир, 1980. – 392 с.
12. Субботин С.А. Синтез нейро-нечетких моделей для выделения и распознавания объектов на сложном фоне по двумерному изображению // Комп'ютерне моделювання та інтелектуальні системи: Збірник наукових праць / За ред. Д.М. Пізи, С.О. Субботіна. – Запоріжжя: ЗНТУ, 2007. – С. 134-146.

С.О. Субботін, А.О. Олійник

Виділення набору інформативних ознак на основі еволюційного пошуку з кластеризацією

Вирішується задача пошуку найбільш інформативної комбінації ознак за допомогою методів еволюційної оптимізації. Запропоновано метод еволюційного пошуку з кластеризацією ознак. Проведено експерименти з виділення інформативного набору ознак для синтезу моделей класифікації автотранспортних засобів.

S.A. Subbotin, A.A. Oleynik

Feature Selection Based on the Evolutionary Search with Clusterization

The problem of the most informative feature combination search based on the evolutionary optimization methods is solved. The method of evolutionary search with clusterization is offered. Experiments on allocation of an informative feature set for synthesis of vehicle classification models are lead.

Статья поступила в редакцию 26.02.2008.